Evaluating large language models as agents in the clinic

Year: 2024 **|** *Citations: 123* **|** *Authors: Nikita Mehandru, Brenda Y. Miao, Eduardo Rodriguez Almaraz*

Abstract

Recent developments in large language models (LLMs) have unlocked opportunities for healthcare, from information synthesis to clinical decision support. These LLMs are not just capable of modeling language, but can also act as intelligent “agents” that interact with stakeholders in open-ended conversations and even influence clinical decision-making. Rather than relying on benchmarks that measure a model’s ability to process clinical data or answer standardized test questions, LLM agents can be modeled in high-fidelity simulations of clinical settings and should be assessed for their impact on clinical workflows. These evaluation frameworks, which we refer to as “Artificial Intelligence Structured Clinical Examinations” (“AI-SCE”), can draw from comparable technologies where machines operate with varying degrees of self-governance, such as self-driving cars, in dynamic environments with multiple stakeholders. Developing these robust, real-world clinical evaluations will be crucial towards deploying LLM agents in medical settings.